MODUL 7

DATA PREPROCESSING

7.1. Tujuan

1. Mahasiswa mampu menyebutkan tipe-tipe data yang digunakan dalam data mining.
2. Mahasiswa mampu menjelaskan permasalahan kualitas data dan penyelesaiannya.
3. Mahasiswa mampu melakukan data preprocessing.

7.2. Landasan Teori

Sebelum diproses data mining sering kali diperlukan preprocessing. Data preprocessing menerangkan tipe-tipe proses yang melaksanakan data mentah untuk mempersiapkan proses prosedur yang lainnya. Tujuan preprosesing dalam data mining adalah mentransformasi data ke suatu format yang prosesnya lebih mudah dan efektif untuk kebutuhan pemakai, dengan indikator sebagai berikut:

1. Mendapatkan hasil yang lebih akurat.
2. Pengurangan waktu komputasi untuk large scale problem.
3. Membuat nilai data menjadi lebih kecil tanpa merubah informasi yang dikandungnya.

Terdapat beberapa alat dan metode yang berbeda yang digunakan untuk preprocessing seperti :

1. *Sampling*, menyeleksi subset representatif dari populasi data yang besar.

©Muhammadiyah university Press

2. *Transformation*, memanipulasi data mentah untuk menghasilkan input tunggal.
3. *Denoising*, menghilangkan noise dari data.
4. *Normalization*, mengorganisasi data untuk pengaksesan yang lebih spesifik.
5. *Feature extraction*, membuka spesifikasi data yang signifikan.

Menurut Susanto (2007), alasan-alasan harus dilakukan data preprocessing adalah antara lain:

1. Data mentah yang ada sebagian besar kotor, seperti:
   a) Tidak lengkap, yaitu berisi data yang hilang / kosong, kekurangan atribut yang sesuai, atau hanya berisi data aggregate.
   b) Banyak noise, yaitu berisi data yang outlier, atau berisi data error.
   c) Tidak konsisten, yaitu berisi nilai yang berbeda dalam suatu kode atau nama.
2. Data yang tidak berkualitas, akan menghasilkan kualitas mining yang tidak baik pula.
3. Data preprocessing, cleaning, dan transformasi merupakan pekerjaan mayoritas dalam aplikasi data mining.

Untuk mengatasi masalah data tersebut, dilakukan preprocessing terhadap data sebelum diolah dengan data mining. Preprocessing dapat dilakukan dengan beberapa teknik yaitu:

1. **Cleaning**, memperkecil jumlah data yang hilang atau berbeda, dapat dilakukan dengan cara:
   a) Mengisi data yang hilang dengan default value.
   b) Mengisi data secara manual, misal: trace ulang transaksi untuk mengetahui data yang hilang.
   c) Mengisi dengan rata-rata atribut tersebut, misal: gaji pegawai yang kosong diisi dengan rata-rata gaji pegawai.

©Muhammadiyah University Press

d) Mengisi dengan rata-rata suatu atribut untuk kelas yang sama, misal: gaji pegawai yang kosong diisi dengan rata-rata gaji pegawai yang memiliki jabatan yang sama.

e) Menggunakan regresi, prediksi berdasarkan dua variabel yang lain, misal: mengisi gaji pegawai yang kosong dengan nilai prediksi dengan regresi berdasarkan jabatan dan lama masa kerja.

f) Menghilangkan baris yang mengandung data yang hilang. misal: menghilangkan data pegawai yang gaji pegawainya kosong.

g) Binning by means, menggunakan rata-rata pengelompokkan. misal: sorted data dibagi menjadi beberapa kelompok, dan dicari rata-rata masing-masing kelompok untuk mengganti setiap data yang ada, sesuai dengan kelompoknya. Misal data dari kelompok A diganti dengana rata-rata kelompok A.

h) Binning by range boundaries, menggunakan batas terdekat suatu kelompok data, misal: sorted data dibagi menjadi beberapa kelompok, dicari nilai minimum dan maximum dari masing-masing kelompok, lalu gantikan tiap nilai di suatu kelompok dengan batas atas atau batas bawah kelompoknya, sesuai dengan yang paling dekat.

i) Mencari dan menghilangkan outlier dengan pengelompokan atau regresi.

j) Binning, mengganti suatu nilai outlier dengan nilai yang lebih sesuai dengan data lain yang ada di sekitar data outlier tersebut (local smoothing).

2. **Integrasi**, menggabungkan beberapa sumber data sehingga dapat saling melengkapi. Data perlu digabungkan dengan key yang sesuai. Key ini mungkin memiliki nama yang berbeda di sumber data yang berbeda. Misal di tabel

©Muhammadiyah University Press

a menggunakan nama atribut 'id', di tabel b menggunakan nama atribut 'nomor', atau satuan yang digunakan untuk konsep yang sama (misal harga) disimpan dalam juta dan ribu.

3. **Transformasi**, mengubah data yang kompleks dengan tidak menghilangkan isi, sehingga lebih mudah diolah, dilakukan dengan cara:
   a) Smoothing (binning, clustering dan regresi).
   b) Agregasi (summarize, menggunakaan dimensi yang lebih general (cube construction)).
   c) Generalisasi, misal menggunakan dimensi propinsi daripada kabupaten atau grouping (hirarki konsep).
   d) Normalisasi, mengelompokkan data sesuai skala tertentu, misal IPK.
      i. Normalisasi min-max, standarisasi data dengan menempatkan data dalam range 0 sampai 1, nilai terkecil sebagai 0, dan nilai terbesar sebagai 1. nilai baru = ((nilai lama – nilai minimal) / (nilai maksimal – nilai minimal)) (range maksimal – range minimal) + range minimal. range minimal = 0, range maksimal = 1.
      ii. Normalisasi z-index, nilai baru = (nilai lama – rata-rata)/standar deviasi.
      iii. Normalisasi skala desimal, nilai baru = nilai lama / 10 ^ x.
4. **Diskretisasi**, membagi nilai data menjadi beberapa range data, dilakukan dengan cara:
   a) Binning, seperti prosedur Cleaning 1.10.
   b) Hirarki konsep, misal mengelompokkan harga produk menjadi, mahal, biasa, murah.
5. **Reduksi**, mengurangi jumlah data sehingga resource yang digunakan lebih sedikit, sehingga prosesnya dapat

©Muhammadiyah University Press

lebih cepat dilakukan dengan cara:

a) Sampling/ generalisasi.

b) Agregasi, seperti agregasi pada transformasi. data ribuan memiliki volume byte yang lebih kecil daripada data jutaan

c) Mengurangi atribut yang tidak perlu (korelasi yang rendah terhadap keseluruhan data), misal nomor telepon, nama ibu atau nama jalan. Jika data set memiliki atribut sejumlah n, maka ada 2^n (2 pangkat n) kemungkinan korelasi antar atribut kompresi data.

7.3. Alat dan Bahan

1. Komputer dengan sistem operasi Windows.
2. Program aplikasi Notepad++ / Notepad.
3. Program aplikasi Weka.
4. Modul Praktikum Data Warehousing dan Data Mining.

7.4. Langkah-langkah Praktikum

7.4.1. Membuat Format Data ARFF (Attribut-Relation File Format)

Teknik data mining dengan menggunakan aplikasi Weka memerlukan sebuah data dalam format ARFF. File ARFF adalah file yang digunakan Weka yang berisi 1 atau lebih *instances* dari atribut.

Hal terpenting dalam data mining jika menggunakan aplikasi Weka adalah membuat file ARFF / membuat instances sebelum diproses klasifikasi, kluster atau hal lainnya. Di dalam format ARFF komponen penting yang harus terdapat dalam file ada 3 macam yaitu:

1. @RELATION, nama_relasi untuk menamakan relasi sebuah file ARFF,

©Muhammadiyah university Press

2. @ATTRIBUTE, untuk penamaan setiap atribut baik numeric, nominal, string dan date.
3. @DATA, terdapat pada setiap line untuk merepresentasikan sebuah instances.

Untuk membuat data dan menyimpan dalam format ARFF bisa dengan menggunakan Notepad/ Notepad++, dengan langkah-langkah berikut ini:

1. Kita ambil sebuah contoh data tentang kemungkinan seseorang akan bermain tenis berdasarkan kondisi cuaca, suhu, kelembaban udara, berangin seperti tabel berikut.

   Nama Tabel : Cuaca

Cuaca	Suhu	Kelembaban_ Udara	Berangin	Bermain_Tenis
Cerah	85	85	TIDAK	TIDAK
Cerah	80	90	YA	TIDAK
Mendung	83	86	TIDAK	YA
Hujan	70	96	TIDAK	YA
Hujan	68	80	TIDAK	YA
Hujan	65	70	YA	TIDAK
Mendung	64	65	YA	YA
Cerah	72	95	TIDAK	TIDAK
Cerah	69	70	TIDAK	YA
Hujan	75	80	TIDAK	YA
Cerah	75	70	YA	YA
Mendung	72	90	YA	YA
Mendung	81	75	TIDAK	YA
Hujan	71	91	YA	TIDAK

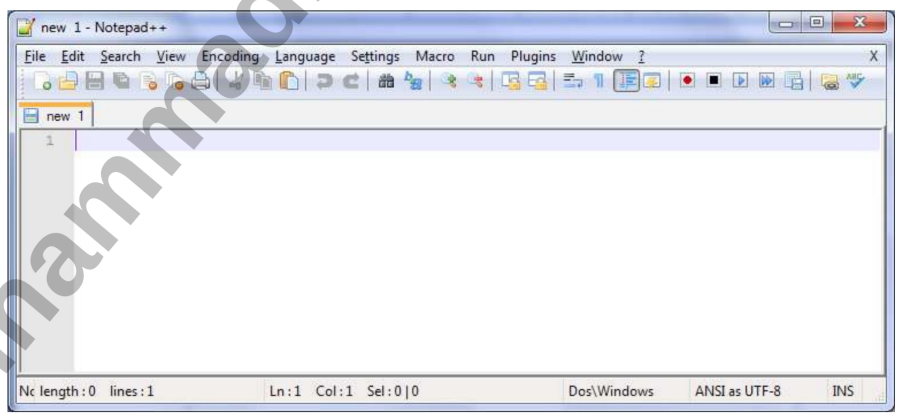
2. Penentuan relasi/ nama tabel. Misalkan kita tentukan nama relasinya adalah Cuaca.
3. Penentuan atribut/ variabel. Berdasarkan tabel cuaca tersebut, tentukan variabel-variabel independen / bebas (X) dan juga variabel dependen / terikat (Y). Variabel independen adalah variabel yang tidak dipengaruhi oleh variabel lainnya. Sedangkan variabel dependen adalah variabel yang nilainya dipengaruhi oleh variabel lainnya.
4. Dalam contoh ini, variabel Bermain_Tenis kita jadikan

©Muhammadiyah University Press

variabel dependen yang akan dicari nilainya berdasarkan kondisi variabel lainnya. Berikut ketentuan variabel / atribut berdasarkan tabel cuaca.

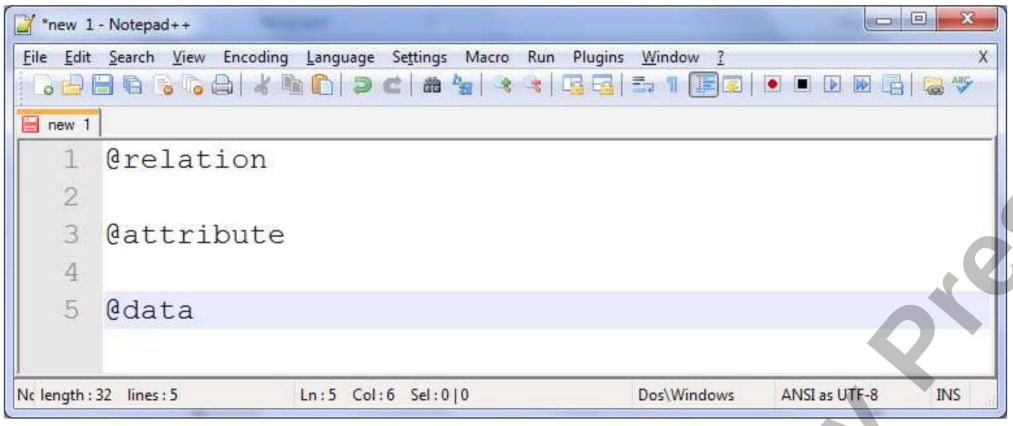
a) Variabel Y = Bermain_Tenis
b) Variabel X1 = Cuaca
c) Variabel X2 = Suhu
d) Variabel X3 = Kelembaban_Udara
e) Variabel X4 = Berangin

5. Penentuan data. Dalam format ARFF, ada 3 jenis data yaitu binomial, polynomial, dan real. Binomial, jika nilai data hanya ada 2 kemungkinan (Contoh: YA dan TIDAK, 0 dan 1, PRIA dan WANITA). Polynomial, jika nilai data lebih dari 2 kemungkinan namun tidak terlalu banyak (Contoh: Cerah, Mendung dan Hujan). Real, jika kemungkinan nilai data sangat beragam dan jumlahnya banyak (Contoh: suhu udara, kecepatan, jarak dan lain-lain) meskipun bisa diringkas menjadi polynomial dengan menggunakan interval sesuai kebutuhan data mining.
6. Buka aplikasi Notepad++ atau Notepad bawaan Windows, cukup ketikkan notepad pada Start Menu.



7. Ketikkan 3 komponen utama dalam format file ARFF yaitu @relation, @attribute, dan @data pada jendela notepad.

©Muhammadiyah University Press



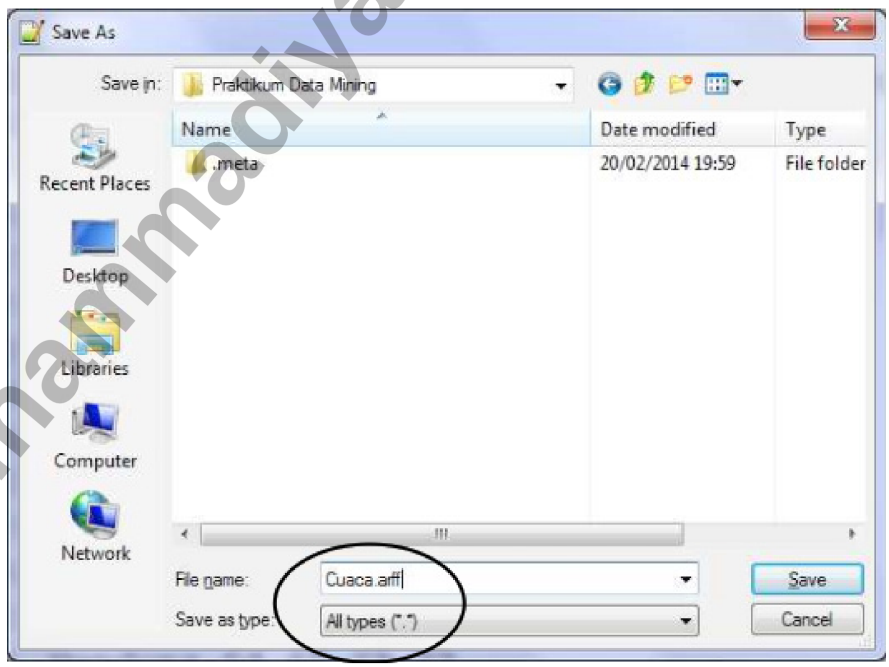
8. Ketikkan nama relasi cuaca di sebelah @relation.
9. Ketik nama atribut dan tipe data masing-masing atribut di sebelah @atribut sesuai jumlah atribut yang digunakan. Jika nilai data bertipe binomial atau polynomial, semua kemungkinan nilai datanya disebutkan dan diletakkan dalam tanda kurung kurawal. Jika nilai data bertipe real, maka cukup tuliskan real.
10. Ketik nilai data di bawah @data untuk tiap baris tabel sesuai dengan urutan atributnya dipisah menggunakan tanda koma. (Catatan: huruf besar dan kecil sangat berpengaruh, sehingga kemungkinan nilai data pada @ attribute harus sama dengan nilai data pada @data).

©Muhammadiyah University Press

```
new 1
1  @relation Cuaca
2
3  @attribute Cuaca {Cerah, Mendung, Hujan}
4  @attribute Suhu real
5  @attribute Kelembaban_Udara ...
6  @attribute ...... {YA, TIDAK}
7  @attribute Bermain_Tenis {YA, TIDAK}
8
9  @data
10 Cerah,85,85,TIDAK,TIDAK
11 Cerah,80,90,YA,TIDAK
12 Mendung,83,86,TIDAK,YA
13 Hujan,70,96,TIDAK,YA
14 .......
15 .......
16 .......
Nc length : 312  lines : 16   Ln : 16  Col : 8  Sel : 0 | 0   Dos\Windows   ANSI as UTF-8   INS
```

Pada titik-titik silakan diisi dan dilanjutkan sesuai dengan **Tabel Cuaca**.

11. Simpan file dengan nama Cuaca.arff. Pastikan untuk memilih **All types (*.*)** pada pilihan **Save as type**.

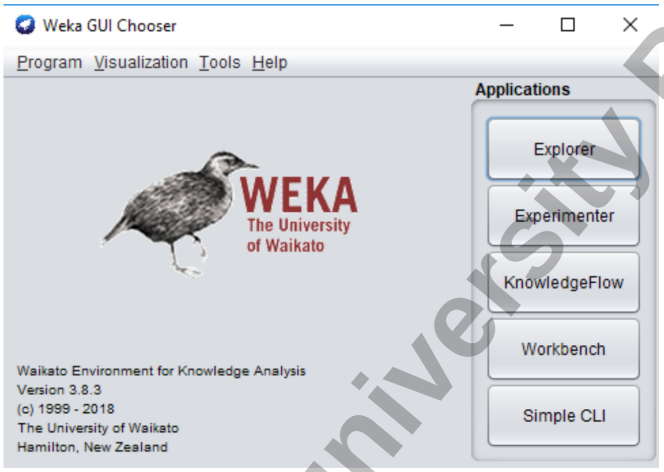


©Muhammadiyah university Press

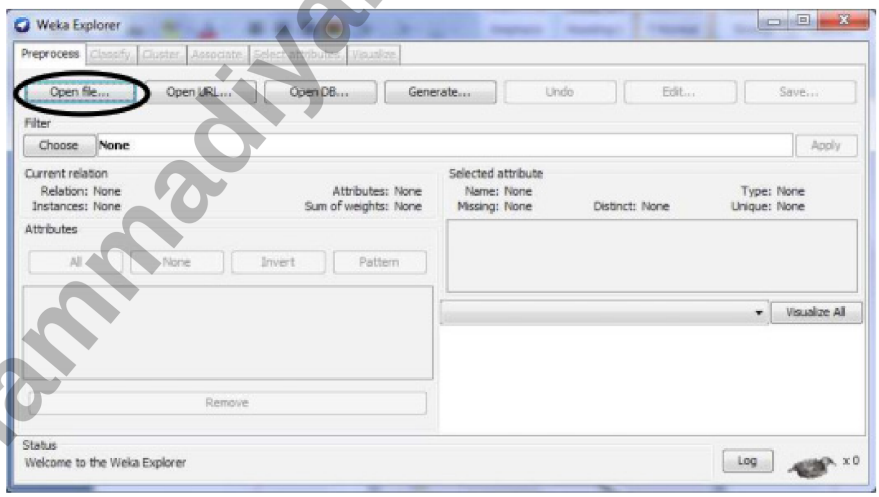
7.4.2. Menggunakan File ARFF dengan Weka

File ARFF hasil dari kegiatan D.1, kemudian akan kita gunakan dalam Weka untuk proses lebih lanjut. Berikut langkah-langkah menggunakan file ARFF dalam Weka:

1. Buka aplikasi Weka.

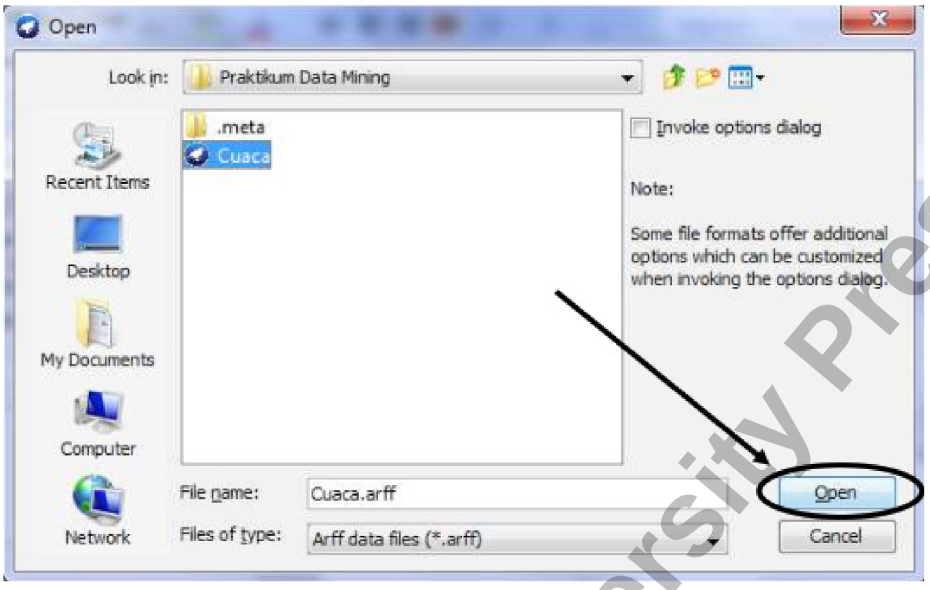


2. Pilih menu **Explorer** sehingga akan muncul jendela Weka Explorer.



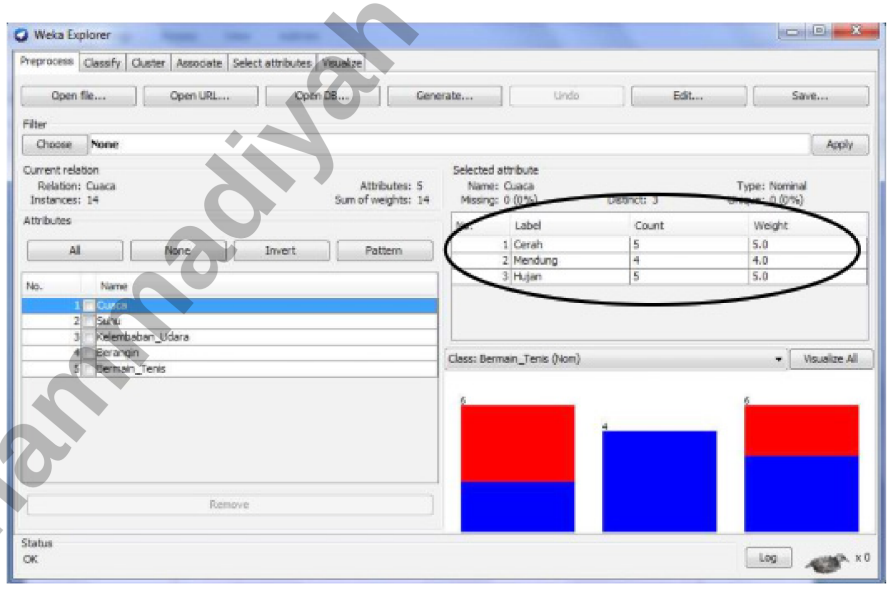
3. Pilih tombol **Open file...** untuk membuka file ARFF yang telah dibuat. Klik **Open**.

©Muhammadiyah University Press



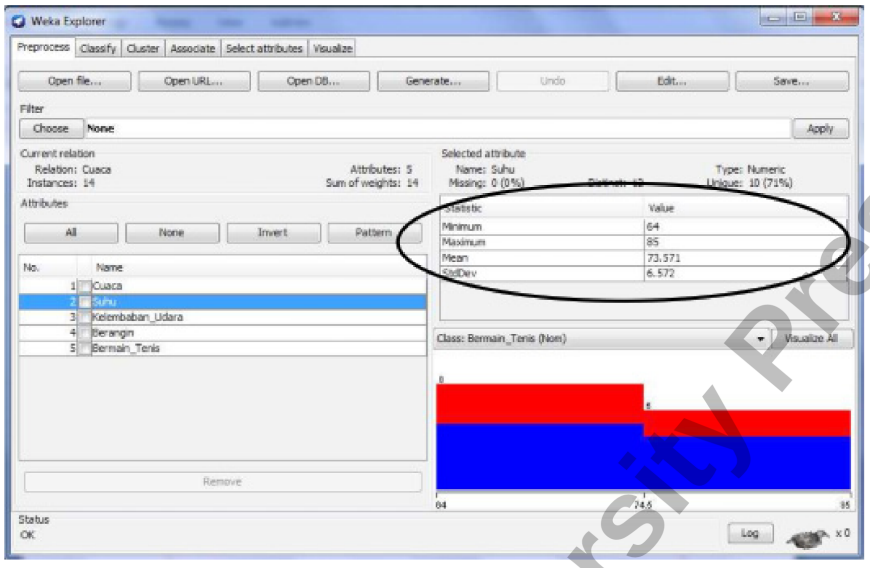
4. Sehingga dalam Weka Explorer akan tampak grafik statistik masing-masing atribut pada Tabel Cuaca.

   Berikut adalah gambar grafik untuk contoh data yang bertipe binomial atau polynomial.



   Berikut ini gambar grafik untuk contoh data yang bertipe real. Mintalah penjelasan kepada dosen atau asisten tentang makna angka-angka dalam lingkaran tersebut.

©Muhammadiyah University Press



7.5. Tugas (Dikerjakan Saat Ini)

1. Buat file ARFF berdasarkan tugas pada Modul 6 soal nomor 1!
2. Perlihatkan file ARFF dan juga gambar grafik untuk setiap data yang ditampilkan dalam Weka berdasarkan file ARFF anda kepada dosen danasisten untuk dinilai!
3. Tulis di atas selembar kertas HVS (mintalah kepada asisten)
4. Berapa jumlah atribut yang bertipe binomial dan polynomial?
5. Berapa jumlah atribut yang bertipe real?
6. Pada atribut Rerata_SKS, berapakah besarnya nilai Maximum, Minimum, Mean, dan StdDev (*Standard Deviation*)?

©Muhammadiyah University Press